Seventh IEEE/ACIS International Conference on Computer and Information Science

Attention Model Based SIFT Keypoints Filtration for Image Retrieval

Ke Gao [1, 2], Shouxun Lin [1], Yongdong Zhang [1], Sheng Tang [1], Huamin Ren [1,3]
[1]*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 100080*
[2]*Graduate University of the Chinese Academy of Sciences, Beijing, China, 100080*
[3]*Beijing University of Chinese Medicine*
{kegao, sxlin, zhyd, ts,renhuamin}@ict.ac.cn

Abstract

Effective feature extraction is a fundamental component of content-based image retrieval. Scale Invariant Feature Transform (SIFT) has been proven to be the most robust local invariant feature descriptor. However, SIFT algorithm generates hundreds of thousands of keypoints per image, and most of them comes from background. This has seriously affected the application of SIFT in real-time image retrieval. This paper addresses this problem and proposes a novel method to filter the SIFT keypoints using attention model. Based on visual attention analysis, all of the keypoints in an image are ranked with their attention saliency, and only the most distinctive keypoints will be reserved. Then we use Bag of words to efficiently index these features. Experiments demonstrate that the attention model based SIFT keypoints filtration algorithm provides significant benefits both in retrieval accuracy and matching speed.

1. Introduction

Local Image Descriptors have been successfully applied in many fields such as object recognition and image retrieval [1], [2]. They are distinctive, robust to occlusion, and do not require segmentation. Recent work has concentrated on making these descriptors invariant to image transformations. There are two considerations to using local image descriptors in these applications. First, the keypoints should be located in position and scale. Typically, these keypoints are placed at local peaks in a scale-space search, thus they are likely to remain stable over transformations. Second, a description of each keypoint must be built, which should be distinctive, concise, and invariant over transformations caused by changes in camera pose and lighting. While the localization and description aspects of keypoints are often interrelated, the solutions to these two problems are independent. Many papers have discussed the second aspect to improve the matching accuracy [3], [4], [5]; while on the contrary, very little work has been done to deal with the background features. Thereby, this paper focuses on the first aspect – the selection of keypoints.

Scale Invariant Feature Transform (SIFT) has been proven to be the most robust among the other local invariant feature descriptors with respect to different geometrical changes [1], [3]. It combines a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. The descriptor is represented by a 3D histogram of gradient locations and orientations, as illustrated in Fig.1 [1]. Recently, PCA-SIFT has been developed based on SIFT algorithm [6]. It applies Principal Components Analysis (PCA) to the normalized image gradient patch, and accelerates matching speed by reducing feature dimensions from 128 to 36 for each patch. However, on a typical image it returns a large number of features, out of which only some fraction lie the object of interest. Especially when the object appears small in the image, the total set of features has a low signal-to-noise ratio. This imposes a great burden on object detectors and image retrieval.

Accordingly, this paper focuses on this problem and proposes a novel method to filter the SIFT keypoints based on attention model. Our contribution lies in proposing a novel method which is well-suited to filter SIFT keypoints. Based on attention model, all keypoints in an image are ranked with its saliency, and only the most distinctive keypoints will be reserved. In this way, the matching speed is accelerated evidently. Moreover, the region information and global image distribution are also taken into account.

978-0-7695-3131-1/08 $25.00 © 2008 IEEE
DOI 10.1109/ICIS.2008.24

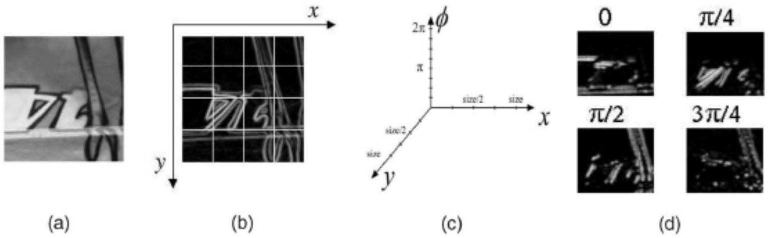


Figure 1. SIFT descriptor. (a) Detected patch. (b) Gradient image and location grid. (c) Dimensions of the histogram. (d) Four of eight orientation planes.

The rest of this paper is organized as follows: Section 2 introduces the relevant aspects of attention model. In section 3, SIFT keypoints filtration using attention model is discussed in detail. Based on the methods presented above, Section 4 provides experimental results in the context of an image retrieval application. Finally, Section 5 summarizes the contributions of this paper.

2. Review of Attention Model

Attention is at the nexus between cognition and perception. While interpreting a complex scene, a human being selects a subset of the available sensory information before further processing. This selection region is so-called “focus of attention”[7]. Visual attention analysis provides an alternative methodology to understand image semantic in many applications, such as adaptive content delivery and region-based image retrieval. A number of computational attention models were developed, such as the models proposed in [7], [8].

Based on these work, Itti proposed a saliency-based attention model for scene analysis [8]. Here “saliency region” means the region which has evident contrast with its surrounding, as shown in Fig. 2.

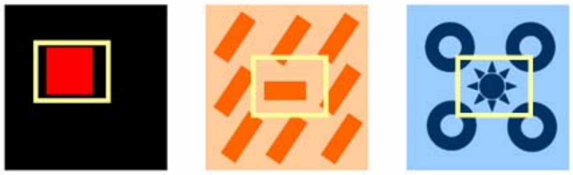


Figure 2. Saliency region based on color, texture and shape perception (as figured out by yellow rectangles in each picture).

In Itti’s work, visual input is first decomposed into a set of feature maps (42 low-level feature maps are extracted separately from different color channels, intensity channels and orientation channels at 8 spatial scales). And then, using a normalization operator, a “saliency map” is generated in a bottom-up manner as a combination of these feature maps. The model is briefly described in Fig 3.

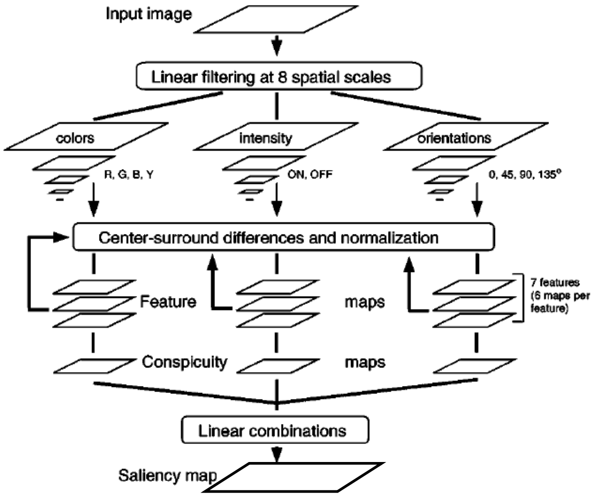


Figure 3. General architecture of the model. The Saliency map is a combination of 42 Low-level visual feature maps.

The saliency map is topology corresponding with the input image, and represents the local “saliency” of each pixel with respect to its neighborhood. The pixel with maximal luminance of this saliency map corresponds to the most salient location of the original image.

3. SIFT Keypoints Filtration using Attention Model

Content-based image retrieval using local invariant can be looked as the problem of transforming the image into a set of feature vectors. For good retrieval performance, the extracted features should satisfy two criteria. The first one is the distinctiveness, which means that the extracted features should distinguish the object image exactly from the other images. The second one is the matching speed. SIFT descriptors are accurate enough, but there are too many keypoints generated from each image, and most of them are "noise points" come from background. Consequently, this paper uses attention model to filter SIFT keypoints. The following is the specific explanation of our novel method.

3.1 SIFT Keypoints Extraction

SIFT, as described in [3], consists of four major stages: (1) scale-space peak selection; (2) keypoints localization; (3) orientation assignment; (4) keypoint descriptor.

In the first stage, potential interest points that are invariant to scale change of the image are identified by scanning over all possible scales and image locations. Because the only possible scale-space kernel is the Gaussian function [3], the scale space of an image is defined as a function $L(x,y,\sigma)$, which is produced from the convolution of a variable-scale Gaussian $G(x,y,\sigma)$, with an input image $I(x,y)$:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y)$$

Where * is the convolution operation in x and y, and

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

To efficiently detect stable keypoint locations in scale space, a series of difference-of-Gaussian (DoG) images are established, because the DoG function provides a close approximation to the scale-normalized Laplacian of Gaussian, $\sigma^2\nabla^2 G$. And the maxima and minima of $\sigma^2\nabla^2 G$ produce the most stable image features among a range of image functions, such as the gradient, Hessian, or Harris corner function.

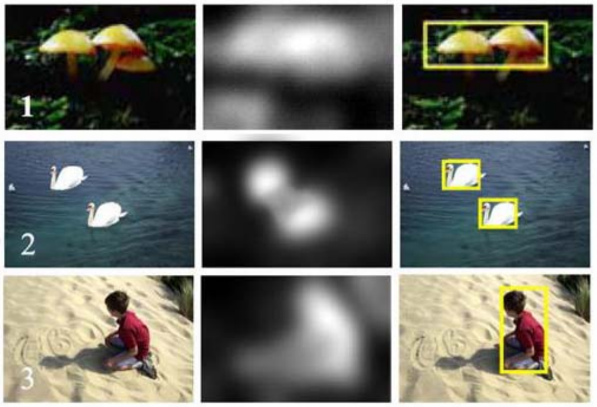
$$\sigma^2\nabla^2 G = \frac{\partial G}{\partial \sigma} \approx \frac{G(x,y,k\sigma) - G(x,y,\sigma)}{k\sigma - \sigma}$$

In the second stage, candidate keypoints are localized to sub-pixel accuracy and eliminated if found to be unstable. The third identifies the dominant orientations for each keypoint based on its local image patch. The final stage builds a local image descriptor for each keypoint, based upon the image gradients in its local neighborhood.

The dimension of standard SIFT descriptor for each keypoint is 128, while the PCA-SIFT reduces the dimension to 36 [6]. Our work is based on the first three stages, and further uses attention model to filter these keypoints, which provides significant benefits both in retrieval accuracy and matching speed.

3.2 Attention Model based Keypoints Filtration

The novel algorithm for SIFT keypoints filtration is based on the first three stages of the standard SIFT descriptor. For each image, after the SIFT keypoints extraction, attention model (described in section 2) is used to generate saliency map. And then, fuzzy growing [9] is performed to find all of the saliency regions for original image. Considering the calculation complexity, the number of saliency regions per image is limited to 3. Fig 4 gives an example in practical application.



(a) (b) (c)

Figure 4. Some sample results of saliency regions detection based on attention model and fuzzy growing. (a) original images, col. (b) attention model based saliency map, col. (c) saliency regions (as figured out by yellow rectangles), col.

As shown in Fig.4, the saliency regions (SR) in saliency map can be in arbitrary shapes. Generally, a SR can be represented by a set of pixels in the original image. However, we use rectangle here for simplicity, and a rectangular SR is defined as {*Center_x, Center_y, Width, Height*}, where (*Center_x, Center_y*) represents the location of the rectangle center, and (*Width, Height*) denotes the size of this rectangular SR. In this paper, we assume that no rectangle will overlap with each other. This simplification accelerates the computation without much information loss.

Based on the definition of SR, each SIFT keypoint is attached with a saliency weight KP_{weight} which is calculated as following functions:

$$KP_{weight} = KPR_{dis} * SR_{weight}$$

$$KPR_{dis} = 1 - \frac{2*\sqrt{(x - Center\_x)^2 + (y - Center\_y)^2}}{\sqrt{Width^2 + Height^2}}$$

$$SR_{weight} = R_{area} * R_{pos}$$

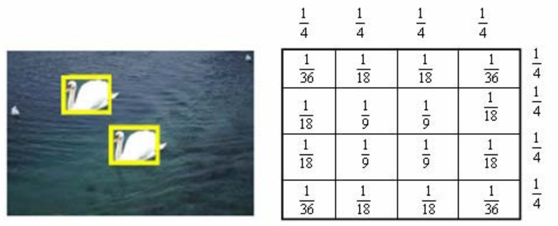
KPR_{dis} is in inverse proportion with the distance between this keypoint and center of the region which contains it. (x, y) denotes the keypoint's location. If the keypoint is in the center of its corresponding region, its KPR_{dis} is 1. Suppose the keypoint isn't subject to any saliency regions detected in this image, its KPR_{dis} is 0.

SR_{weight} denotes the saliency weight of this saliency region. We observe that the importance of a detected region is usually reflected by its region area weight R_{area} and position weight R_{pos}.

If a region is too small to provide any useful information, it would not be considered. Therefore, only the regions bigger than 5% of total image are ranked with their area, and only the top 3 regions will be reserved as SRs. Suppose an image contains n SRs (n is between 0 and 3), and $area_i$ is the area of each SR. Position weight R_{area} of the current SR is calculated as the following function:

$$R_{area} = \frac{area_{current}}{\sum_{i=1}^{n} area_i}$$

Since people often pay more attention to the region near the image center, a normalized Gaussian template centered at the image is used to assign the position weight R_{pos}, which is determined by the center position of the current SR. As shown in Fig.5, The summation of all position weight in an image is 1.



$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	
$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	$\frac{1}{4}$
$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{4}$
$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{4}$
$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	$\frac{1}{4}$

(a) (b)

Figure 5. Gaussian template of position weight. (a) An example of SR detection. (b) The position weight template.

As discussed above, the saliency weight KP_{weight} of each SIFT keypoint is generated. We rank all keypoints in an image with their KP_{weight}, and only the top N keypoints will be reserved to extract SIFT descriptors. N is determined by the practical application to achieve appropriate balance between retrieval accuracy and speed. In this way, SIFT Keypoints Filtration using Attention Model is accomplished.

The experimental results are provided in detail as following section.

4. Experiment Evaluation

We evaluate the performance of our novel method on a real image data set which consists of three categories: (1) The same object with different background or under different viewpoints. (2) Video frames extracted from some movies. (3) Usual images with different size and content. Most of the original photos are downloaded from the famous ALOI (http://staff.science.uva.nl/~aloi/) and the Caltech gallery (http://vision.caltech.edu/Image_Datasets/Caltech256/). Some samples are shown in Fig.6.

Some geometric and photometric transformations have been made to evaluate the algorithm under different conditions. According to different objects, the data set is divided into about 50 classes, and each class has more than 20 relevant images. There are nearly 6,000 images and 7,240,000 standard SIFT keypoints in all which have been extracted from the image data set.

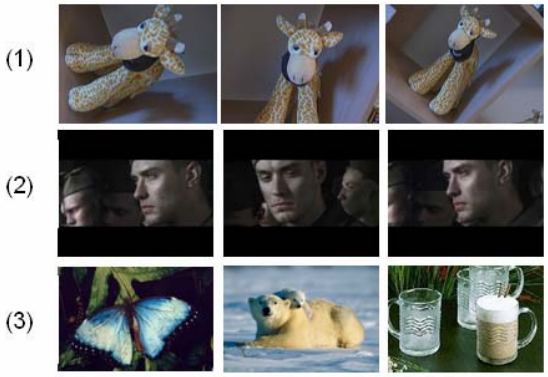


Figure. 6. Example images of the three categories.

4.1 Evaluation Metrics

For image matching, we use the famous method Bag of Words proposed in [10], which vector quantizes the SIFT descriptors into clusters uses k-means, and then represents an image as a bag of "words". Using 'term frequency' as standard weighting, all of the images are organized as an inverted file, and image matching is based on cosine between these quantized vectors. This method can ensure in-time retrieval, and proven to be very useful.

If the cosines distance between image vectors larger than the chosen threshold, this pair of images is called a *match*, and all of the images will be ranked with the matching degree.

To describe the image ranking sequence of image retrieval in this data set, we adopt average retrieval precision. The precision of top n images is calculated as function (8) and (9). Here q is the query image, p_i denotes each image of ranking result, and n is 20.

$$AP_q = \frac{1}{n}\left(\sum_{i=1}^{n}\left(\frac{1}{i}\sum_{j=1}^{i}\psi(p_i,q)\right)\right)$$

$$\psi(p_i,q) = \begin{cases} 1, & \text{if } p_i \text{ is relevant to q} \\ 0, & \text{if } p_i \text{ is not relevant} \end{cases}$$

4.2 Experimental Results and Discussion

The experiment presents results comparing our attention model based SIFT keypoints filtration algorithm (AF-SIFT) to the standard SIFT and PCA-SIFT. The standard SIFT descriptor is created by building smoothed orientation histograms to describe the patch around the keypoint. A 4*4 array of histograms, each with 8 orientation bins was calculated, thus the dimension of standard SIFT is 128. PCA-SIFT descriptor dimension for each keypoint is 36. As to AF-SIFT, we use two methods to compare its performance. AF-SIFT1 uses 128-dimension descriptors in the standard way, while AF-SIFT2 uses a 2*2 array with 8 orientation bins, and its dimension is 32. Both the number of filtered feature and the retrieval accuracy are taken into account, and the experiment is accomplished on a 1GHz Pentium 4 processor.

Our initial goal was to explore more effective alternatives and to empirically evaluate the tradeoffs, so the number of filtered features in this image data set is first presented.

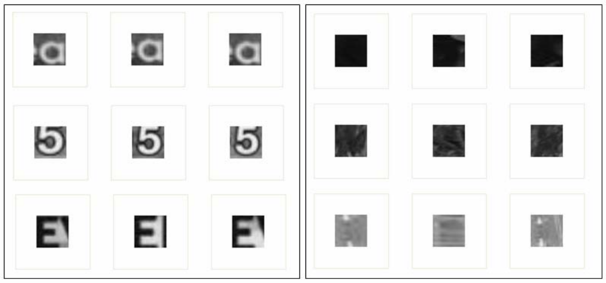

(a) (b)

Figure 7. (a) Reserved features after filtration and clustering. Each row means a "word" which describes a cluster of similar features. (b) Example of removed features.

We can see clearly that our filtration algorithm could effectively remove the features which come from background, and most of them have few information. The process of filtration can effectively avoid interference from these background "noise patch", and improve the quality of k-means cluster.

Table 1. Filtering probability within different image dataset.

Image Dataset	Image amount	SIFT amount	Filtering probability
ALOI	0.5K	175,115	7.4%
Movie Frames	2.5K	1150,573	5.9%
Coral Gallery	3K	1690,812	6.1%

The above table shows the extraction feature amount and filtering probability for each image dataset. It's a bit time-consuming for the series of filtering algorithms, but the processing is completed off-line, and it could effectively reduce the background features, so it in fact decreases the whole calculation time. The SIFT amounts shown above are the original number of features before filtering procession.

Due to the difference of the three image dataset, the filtering probabilities are also not the same. Images from ALOI have few background confusion, they emphasize the infection of different condition, such as varying illumination or view point; on the contrary, frames extracted from movie are almost in the same condition, so their content are much more similar, but there are little obvious difference between foreground and background. As to the Coral Gallery, they are nature photos with confusion background, and the resolution of these images are the biggest in the three dataset, so they have the maximum SIFT feature amount.

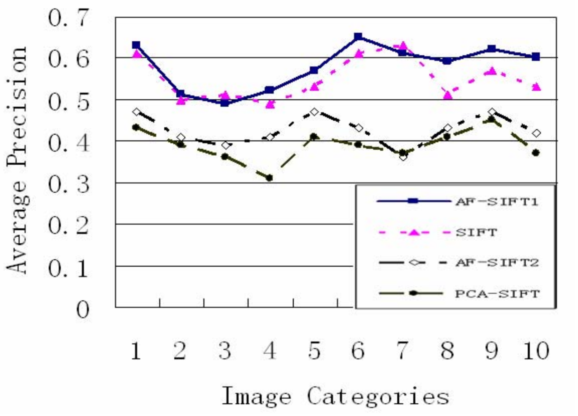


Figure 8. Evaluation for different algorithms.

Afterwards, Fig 8 compares the retrieval accuracy of these four algorithms in the whole image dataset. As shown in Fig8, in most of the tests, AF-SIFT1 obtains the best results, followed by original SIFT. AF-SIFT2 and have lower accuracy, closely followed by PCA-SIFT. It demonstrates that our novel method provides an effective alternative of standard SIFT, and is more appropriate for image retrieval.

The reason of AF-SIFT performance maybe as follows: it filtrates SIFT keypoints based on attention model, which provides information of saliency regions. The area and position of each saliency region are also taken into account. So the ranking of keypionts is based on the global distribution, not only relies on local patches. Then the most distinctive keypoints are reserved, it could effectively avoid the infection of background features, and made the cluster result become more exact. Therefore retrieval accuracy can be improved.

Fig 9 shows how the matching reliability varies as a function of N. Here N denotes the number of SIFT keypoints left behind the filtration, which changes with the threshold predefined in the filtration process. It is easy to presume that increasing the number of SIFT keypoints will result in better accuracy, since the representation is able to capture the structure of the gradient patch with better fidelity. We obtain good results when N equals to 300. Therefore, a good tradeoff between accuracy and speed should be achieved in practical application.

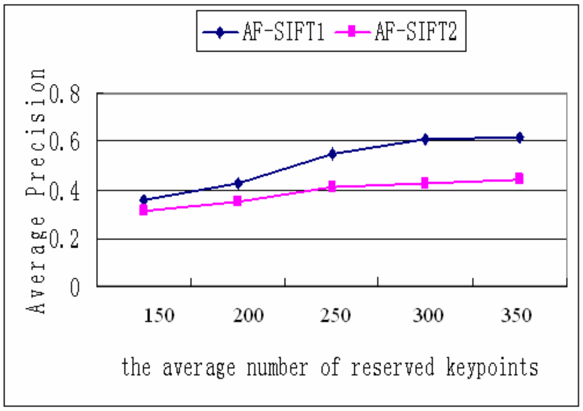


Figure 9. AF-SIFT performance as the number of reserved keypoints is varied.

5. Conclusion

This paper introduced a novel method for SIFT keypoints filtration based on attention model (AF-SIFT). Based on visual attention analysis, all keypoints in an image are ranked with its saliency weight, and only the most distinctive keypoints will be reserved. In this way, the background features can be reduced evidently, and the retrieval accuracy can be improved at the same time. Compared to other local image descriptors, AF-SIFT provides an effective alternative of standard SIFT, and is more appropriate for image retrieval. Of course, this method can also be used with other affine covariant region such as MS, SA and SURF. We are currently extending our algorithm to region-based image retrieval, and seeking for ways to apply this idea to large image database retrieval.

6. Acknowledgement

This work was supported in part by the National Basic Research Program of China (973 Program, 2007CB311100) and the National Nature Science Foundation of China (60773056), the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), the Beijing New Star Project on Science & Technology (2007B071).

7. References

[1] Mikolajczyk K, Schmid, C, "*A Performance Evaluation of Local Descriptors"*. IEEE Trans.Pattern Analysis and Machine Intelligence, 2005, 27(10), p1615-1630

[2] V. Ferrari, T. Tuytelaars, and L. Van Gool. "*Simultaneous Object Recognition and Segmentation by Image Exploration"*, Proc. Eighth European Conf. Computer Vision, 2004, p40-54

[3] D. Lowe, "*Distinctive Image Features from Scale-Invariant Keypoints"*, Int'l J. Computer Vision, vol. 2, no. 60, 2004, p91-110

[4] Abdel-Hakim AE, Farag AA, "*CSIFT: A SIFT Descriptor with Color Invariant Characteristics"*. Computer Vision and Pattern Recognition, 2006,Vol. 2, p1978-1983

[5] T. Tuytelaars and L. Van Gool, "*Matching Widely Separated Views Based on Affine Invariant Regions"*, Int'l J. Computer Vision, 2004,Vol. 1, no. 59, p61-85

[6] Yan Ke, Rahul Sukthankar, "*PCA-SIFT: A More Distinctive Representation for Local Image Descriptors"*. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2004,Vol.2, p506-513

[7] J. K. Tsotsos, S. M. Culhane, W.Y.K. Wai, et al, "*Modeling visual attention via selective tuning"*, Artificial Intelligence, 1995,78: p507-545

[8] Itti L, Gold C, Koch C, "*Visual attention and target detection in cluttered natural scenes"*. Optical Engineering, 2001,40(9), p1784-1 793

[9] Ma Y F, Zhang H J, "*Contrast-based image attention analysis by using fuzzy growing"*. Proceedings of the 11th ACM International Conference on Multimedia. Berkeley, CA, USA: ACM, 2003, p374 – 381

[10] J.Sivic, A.Zisserman, "*Video Google: A Text Retrieval Approach to Object Matching in Videos"*. Proceedings of the International Conference on Computer Vision, 2003, p1470-1477